Nach Data Warehousing kommt Business Intelligence

Andrea Kennel
Trivadis AG
Glattbrugg, Schweiz

Schlüsselworte:
Business Intelligence, Data Warehouse

Zusammenfassung
Data Warehouse bedeutet, dass operative Daten über längere Zeit gesammelt und dann ausgewertet werden. Business Intelligence sucht in vorhandenen Daten business relevante Information. Wo liegt nun der Unterschied? Der Unterschied liegt vor allem in der Betrachtungsweise. Data Warehouse geht von den vorhandenen Daten, Business Intelligence von der business-relevanten Information aus. Business Intelligence ist das Ziel, Data Warehouse der Weg dazu.
Dieser Artikel beschreibt, wie in einem Data Warehouse Projekt und in einem Business Intelligence Projekt vorgegangen werden kann und wo die zentralen Probleme und Stolperfallen liegen.

Was ist BI?

Es geht bei Business Intelligence darum, aus Daten wirtschaftlich relevante Erkenntnisse zu gewinnen. Dazu gehören aber nicht nur Daten, wie sie in einem Data Warehouse (DWH) gespeichert sind. Es gehört vor allem auch die Auswertung dieser Daten dazu, egal ob man nun betriebswirtschaftliche Kennzahlen berechnet oder unbekannte Zusammenhänge mit Data Mining herausfindet.

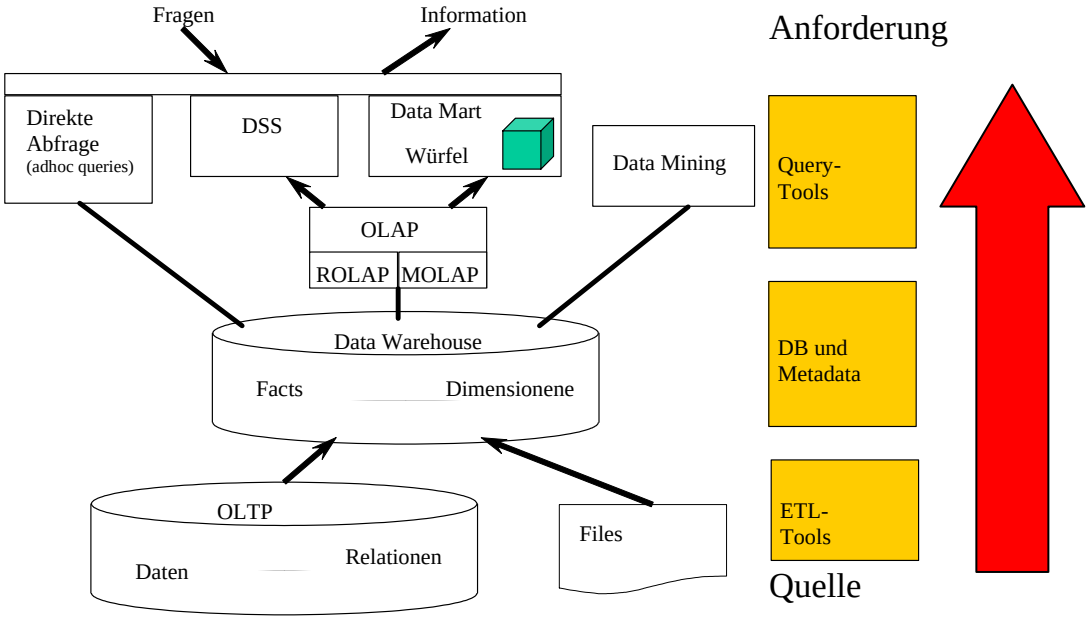
BI ist an Information interessiert, die für das Business wichtig sind. Welche Fülle von Informationen aus den Daten extrahiert werden kann, ist nicht zentral. Zentral ist, dass die Information zur Verfügung gestellt wird, die fachlich zur Führung und Kontrolle des Business relevant ist. Daten sind nur die Grundlage zur Information, aus der Wissen extrahiert werden kann.

Von der Quelle zu den Anforderungen

Das DWH hat hier eine andere Betrachtungsweise. Im DWH sind die Daten der zentrale Punkt. Je nachdem, wie die Daten gesammelt und ausgewertet werden, kann mehr Information gewonnen werden. Und doch kommt BI nicht ohne DWH aus.

> Im Data Warehouse werden Daten gesammelt. Erst wenn diese sinnvoll genutzt werden, entsteht ein **Mehrwert**.

Wie erreichen wir dieses Ziel von Mehrwert?



Grafik: Grundstruktur eines DWH

Ein Data Warehouse sammelt Daten aus unterschiedlichen Quellen und speichert diese über die Zeit. So sind Auswertungen über Zeiträume und unterschiedliche Datenquellen möglich. Zur Auswertung gibt es verschiedene Möglichkeiten von direkten Abfragen über OLAP bis hin zu fertigen Reports oder Data Mining.

Eines der Probleme im DWH ist das Zusammenführen der unterschiedlichen Quellen. Dabei muss festgestellt werden, was in welchen Quellen vorhanden ist, was verschieden benennt ist und dasselbe bedeutet, oder was gleich benannt ist und doch unterschiedliche Bedeutung hat. So werden die Quellen von Entwicklern oder DWH-Architekten zusammen mit den Entwicklern der Quellsysteme analysiert. Dabei wird auch festgelegt, wie die unterschiedlichen Daten konsolidiert werden können. Der nächste Schritt ist die Planung der Historisierung. Die Daten im Warehouse werden über längere Zeit gespeichert. So muss festgelegt werden, wie mit Änderungen der Daten umgegangen wird. Dabei muss unter anderem festgelegt werden, welche Dimensionen versioniert werden und bei welchen sicher nur der aktuelle Stand interessiert.
Basierend auf diesen Abklärungen über der Quellen, Konsolidierung und Historisierung wird dann das Datenmodell des DWH erstellt. Nun kommt die aufwändigste Phase: die Umsetzung des Datenmodells und das Schreiben oder Definieren der Ladeprozesse. Damit diese Phase nicht zu lange geht, wird oft nur ein erster Teil implementiert. Dazu wird festgelegt, welche Daten für erste Auswertungen gebraucht werden. Sind diese Daten geladen, werden die ersten Auswertungen erstellt und ausgeliefert. Dieser Schritt ist für die Akzeptanz des ganzen DWH wichtig. Bekommt der Endbenutzer als erstes Auswertungen, die für ihn wichtig sind und einen Mehrwert darstellen, so hat das DWH grosse Chancen weiter entwickelt zu werden. Sind diese ersten Auswertungen nichts sagend oder verwirrend, sinkt die Akzeptanz. Hier hat der DWH Spezialist also eine heikle Aufgabe zu lösen. Diese kann sicher nicht ohne Einbezug des Endbenutzers gelöst werden.

Axiom 1
Läuft das DWH Projekt gut, kommt jetzt die Iterative Phase. Es werden weitere Quelldaten geladen und es werden weitere Auswertungen zur Verfügung gestellt. Dies lässt sich in folgendem Axiom zusammenfassen:

> Mit dem Essen kommt der Hunger

… oder …

> Wird ein Data Warehouse genutzt, so entstehen neuen Anforderungen

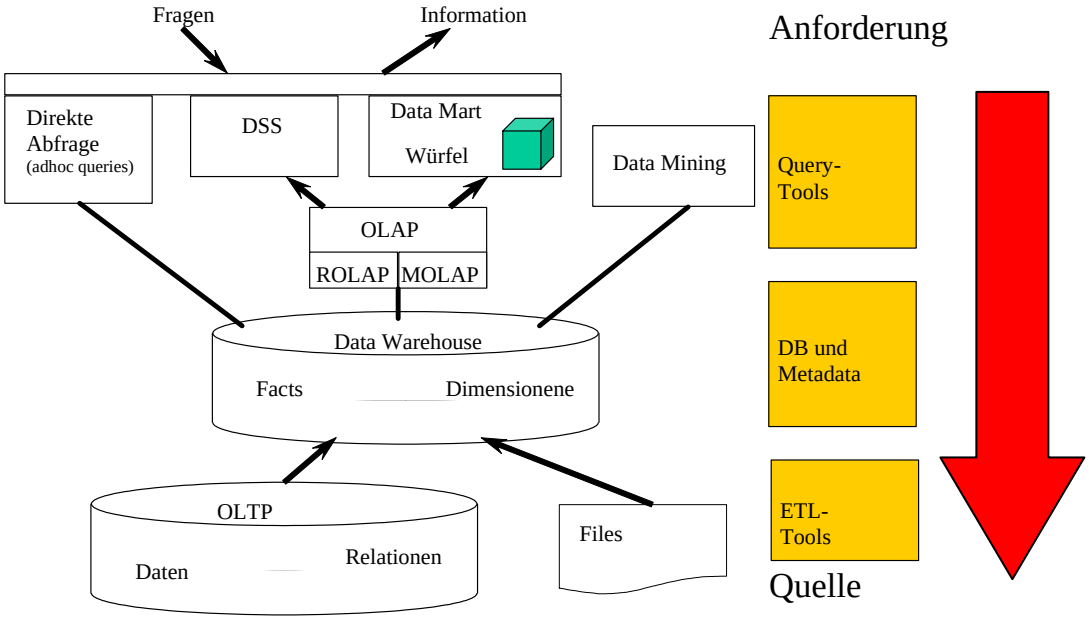
… oder …

> Entstehen neue Anforderungen, so wird ein Data Warehouse genutzt

Von den Anforderungen zur Quelle
Betrachten wir nochmals zusammengefasst, wie ein DWH aufgebaut ist. Der Entwickler und der DWH-Architekt analysieren die Quellsysteme. Daraus wird definiert, wie das DWH aufgebaut wird. Basierend auf den vorhandenen Daten wird definiert, welche Auswertungen möglich und für das Business interessant sein könnten. Der Entwickler zeigt dem Endbenutzer, was mit den Daten möglich ist. Die Daten sind der zentrale Punkt, aus ihnen entsteht Information.
Am Anfang dieses Artikels wurde festgehalten:

> Im Data Warehouse werden Daten gesammelt. Erst wenn diese sinnvoll genutzt werden, entsteht ein **Mehrwert**.

Nun stellt sich die Frage, ob mit dem gewählten Vorgehen der Mehrwert wirklich sichergestellt werden kann. Die Information hängt nicht davon ab, was der Endbenutzer primär will, sondern davon, was das System liefern kann. Sinnvoller wäre und ist es, von den Anforderungen aus zu gehen. Zentral sind nicht die Daten, sondern die Anforderungen und damit das Informationsbedürfnis des Endbenutzers.



Der Endbenutzer legt fest, welche Information er für seine Arbeit braucht. Dazu müssen Fragen zum Business gestellt werden. Hier eine Auswahl möglicher Fragen:

Wie wird Erfolg definiert?
Von was hängt der Bonus ab?
Welche Kennzahlen definieren der Erfolg?
Wie und wie oft wird der Erfolg gemessen?
Wie werden Probleme identifiziert?
Welches sind heute bekannte Probleme?
Welche Information braucht es, um Ziele besser zu erreichen?

Diese Fragen müssen vom Endbenutzer beantwortet werden und dem DWH-Architekten erklärt werden. Gemeinsam muss nun festgelegt werden, welche Daten die benötigte Information liefern. So muss festgestellt werden, welche Anforderungen mit welchen Quelldaten abgedeckt werden können. Dabei ist immer wieder erstaunlich, dass der Endbenutzer die Quellsysteme recht gut kennt. Er kann nicht sagen in welcher Tabelle in welchem Attribut die benötigten Daten stehen, er weiss aber, in welchem System die Quelldaten gepflegt werden und in welcher Qualität diese vorhanden sind. Dieses Wissen hilft bei der Quellanalyse und vereinfacht diese.
Nach dieser ersten gemeinsamen Phase der Analyse muss nun festgelegt werden, welche Quelldaten benötigt werden und welche nie gebraucht werden. So entsteht ein Datenmodell für das DWH, das sich nach den Auswertungen und nicht nach den Quellen richtet. Der Nachteil dabei ist klar, dass die Ladeprozeduren komplexer werden. Der Vorteil dagegen liegt bei den Abfragen die einfacher sind und dabei, dass nur die Quellen, die für das Business auch wirklich relevant sind, geladen werden müssen. Das DWH wird so etwas schlanker, einfacher und schneller.

Axiom 2
Bei diesem Vorgehen stehen klar nicht die Daten im Vordergrund, sondern die benötigte Information. Daraus lässt sich ein zweites Axiom herleiten:

- Der Kunde ist an Information, nicht am DWH interessiert

… oder …

- Das Ziel ist wichtig, nicht der Weg

Dieses zweite Axiom widerspricht dem ersten Axiom aber in keiner Weise. Im Gegenteil, das erste Axiom, dass mit dem Essen der Hunger kommt, bleibt. Wird die Information in das Zentrum gestellt, so sind die ersten Auswertungen, die der Endbenutzer erhält sicher interessant. Schnell kommt das Bedürfnis nach neuen Informationen mit anderen Kombinationen von Basisdaten.
Bei der Architektur des DWH muss dies berücksichtigt werden. Die Architektur muss so gewählt werden, dass auch neue Anforderungen möglichst einfach abgedeckt werden können. Konkret heisst das, dass auch Daten, die für die ersten Auswertungen nicht wichtig sind geladen werden sollten oder die Daten detaillierter geladen werden, als im ersten Schritt gebraucht. Ein gutes, flexibles DWH-Design bedingt, dass der DWH-Architekt sich intensiv mit den Businessanforderungen auseinander setzt und sich in das Business hineindenken kann. Ein guter DWH-Architekt versteht nicht nur etwas von Datenbanken, sondern auch

etwas von der Anwendung. So wie ein guter Architekt selber schon gekocht haben sollte oder mit einem Koch sprechen sollte, bevor er eine Küche zeichnet.

Wir haben gesehen, dass bei DWH von der Quelle zur Auswertung entwickelt wird und bei BI genau umgekehrt. So gesehen ist BI die Umkehrung von DWH. Oder anders gesagt: BI geht vom Business aus, DWH von den Daten.

Fachliche Probleme
Doch auch bei BI bleiben diverse Probleme, die wir von DWH kennen. Diese Probleme können grob in fachliche und technische Probleme unterteilt werden. Wir betrachten zuerst die fachlichen Probleme.

Der Endbenutzer spricht oft nicht dieselbe **Sprache** wie ein Informatiker. Seine Anforderungen sind eher vage und mathematisch gesehen nicht sehr präzise. So wird oft gesagt, dass man die Zahlen auf Wochen und dann auf Monate verdichtet sehen möchte. Der Informatiker sieht hier sofort, dass eine Woche nicht zwingend zu genau einem Monat gehört. Hier muss der Informatiker immer wieder nachfragen, um Antworten zu präzisieren und konkretisieren. Idealerweise führt der Informatiker mehrere Interviews mit dem Endbenutzer. Diese Interviews können unterstützt werden durch grafische Darstellung der Zusammenhänge (ADAPT) und durch tabellarische Beispiele mit Excel. Wichtig ist, dass definiert werden kann, welche Fakten und Dimensionen interessieren. Dies aber auf einer abstrakten logischen Ebene.
Beschlüsse über logisches Datenmodell, Granularität und Historisierung sollen mit den daraus folgenden Konsequenzen protokolliert werden.

Eine grosse Quelle für Missverständnisse sind **Fachbegriffe**, die vermeintlich klar sind. So ist für den Informatiker klar, dass DB2 eine Datenbank ist. Für den Controller aber ist DB2 der Deckungsbeitrag 2, von dem er genau sagen kann, wie dieser berechnet wird. Frage ich den Projektleiter, wie viele Personen im Projekt mitarbeiten, so sagt er mir 4 Personen. Frage ich den Abteilungsleiter, so sagt er mir 240%. Beides stimmt, es kommt darauf an, ob ich die Personen oder die Stellenprozente meine.
Diese Beispiele sollen zeigen, dass je nach Kontext oder Sichtweise dieselben Begriffe unterschiedliche Bedeutung haben können. Da ist es wichtig, ein Glossar zu erstellen, in dem genau festgelegt ist, welcher Begriff was genau bedeutet. Begriffe, die mehrere Bedeutungen haben, sollten durch präzisere Begriffe ersetzt werden. So sollte nicht von Personen, sondern von Personenzahl (Headcount) oder Stellenprozenten gesprochen werden. Für Kennzahlen, die berechnet sind, muss die entsprechende Formel angegeben werden. Bei Kennzahlen ist auch wichtig, dass möglichst früh festgehalten wird, über welche Dimension diese aufsummiert werden kann.

Technische Probleme
Bis hierhin haben wir uns mit Themen beschäftigt, die nicht direkt mit Oracle zu tun haben. Das Vorgehen und die Gespräche mit dem Endbenutzer hängen nicht von der gewählten Datenbank ab. Die technischen Probleme und technischen Möglichkeiten jedoch schon.

Zu den technischen Problemen gehört sicher die konkrete **physische Modellierung**. Es muss festgelegt werden, ob mit Star-Schema oder mit Analytic Workspace gearbeitet werden soll. Weiter muss festgelegt werden, welche Tabellen wie partitioniert werden können. Für die Performance der Abfragen sind weiter der Denormalisierungsgrad und die Indexierung wichtig.
Für **ETL** muss überlegt werden, welches Tool zum Einsatz kommt, und wie die Performance optimiert werden kann. Oft erlauben Tools relativ einfache grafische Entwicklung, doch ist dann die Optimierung nicht immer einfach. Bei zeitkritischen Prozessen lohnt es sich oft, eine PL/SQL-Prozedur mit Bulk einzusetzen und diese Prozedur dann dem Tool bekannt zu machen.
Ein zentraler Punkt ist die Umsetzung der **Dimensionen**. Dabei muss festgelegt werden, welche Dimension Versionen enthalten soll. Dies ist normalerweise durch die Anforderungen bereits gegeben. Beim Umsetzen der Dimensionen läuft man Gefahr, viele sehr kleine Dimensionen oder aber einzelne sehr grosse Dimensionen zu erhalten. So genannte „Fast changing monster dimensions“ sind zu vermeiden. Dieses Problem kann entstehen, wenn bei einer grossen Dimension wie Kunde auch eigentlich Fakten wie der Kontostand abgelegt wird. In einem solchen Fall muss die Aufteilung zwischen Fakten und Dimensionen überarbeitet werden. Doch auch wenn keine eigentlichen Fakten in den Dimensionen vorkommen, können diese häufig ändern und zu gross werden. Detaillierte Ausführungen, wie dies vermieden werden kann, würde den Rahmen dieses Vortrages sprengen. In diesem Konferenzband ist dazu ein Artikel von Karol Hajdu mit dem Titel „Lange Antwortzeiten bei grossen Datamarts? Dies muss nicht so sein!“ zu finden.

Schlussfolgerung
Business Intelligence ist eigentlich die logische Folge von Data Warehousing. Waren ursprünglich die Daten im Zentrum, so ist nun das Business und damit die Information ins Zentrum gerückt. Daten sind nicht mehr das Ziel, sondern das Mittel zum Zweck. Der Endbenutzer will nicht primär ein Data Warehouse, er will Informationen. Die Daten und die bekannten Problemstellungen des Data Warehousings bleiben, sollen den Endbenutzer aber nicht beschäftigen. Damit der Endbenutzer das bekommt, was er für seinen Job braucht, braucht es im BI Umfeld verschiedenes Fachwissen. Es braucht einerseits das Business Fachwissen, dann technisches Wissen über Data Warehousing und natürlich weiterhin Grundwissen über Datenbanken.
Erst wenn Fachleute aus diesen unterschiedlichen Gebieten gut zusammen arbeiten, entstehen Systeme, die den Namen Business Intelligence verdient haben.

Literatur

Generelle Grundlagen englisch

The Data Warehouse Toolkit, R. Kimball. Wiley Computer Publishing, ISBN 0-471-15337-0

The Data Warehouse Lifecycle Toolkit, R. Kimball. Wiley Computer Publishing, ISBN 0-471-25547-5

Generelle Grundlagen deutsch

Data Warehouse Systeme, A. Bauer, H. Günzel. Dpunkt.verlag, ISBN 3-932588-76-2

Data Warehousing und Data Mining, M. Lusti. Springer, ISBN 3-540-42677-9

Oracle spezifisch

Oracle9iR2 Data Warehousing, Digital Press, ISBN 1-55558-287-7

Dimensionale Modellierung

„Lange Antwortzeiten bei grossen Datamarts? Dies muss nicht so sein!“, Karol Hajdu, in diesem Konferenzband

Kontaktadresse:

Andrea Kennel
Trivadis AG
Europa-Strasse 5
CH-8152 Glattbrugg

Telefon: +41(0)1-808 70 20
Fax: +49(0)1-808 70 21
E-Mail andrea.kennel@trivadis.com
Internet: www.trivadis.com